Bio-inspired Human Action Recognition With A Micro-Doppler Sonar System

Thomas S. Murray, *Member, IEEE,* Daniel R. Mendat, *Member, IEEE,* Kayode A. Sanni *Member, IEEE,* Philippe O. Pouliquen *Member, IEEE,* and Andreas G. Andreou, *Fellow, IEEE*

***Abstract*—This paper explores computational methods to address the problem of doing inference from data in multiple modalities where there exists large amounts of low dimensional data complementary to a much smaller set of high dimensional data. In this instance the low dimensional time-series data are active acoustics from a bio-inspired micro-Doppler sonar sensor system that include no or very limited spatial information, and the high dimensional data are RGB-Depth data from a 3D point cloud sensor. The task is human action recognition from the active acoustic data. To accomplish this, statistical models, trained simultaneously on both the micro-Doppler modulations induced by human actions and symbolic representations of skeletal poses, derived from the 3D point cloud data, are developed. This simultaneous training enables the model to learn relations between the rich temporal structure of the micro-Doppler modulations and the high-dimensional pose sequences of human action. During runtime, the model relies purely on the active acoustic sonar data to infer the human action. Our approach is applicable to other sensing modalities such as the millimeter wave electromagnetic radar devices.**

***Index Terms*—active acoustics, human action recognition, micro-Doppler effect, multimodal action dataset, multistatic sonar, micro-Doppler modulations**

I. Introduction

HUMAN actions range from simple motions, such as a hand wave, to complex sequences composed of many intermediate actions, such as figure skating. Every day each of us performs many actions, even creating new actions to accomplish a novel task. Moreover, we are able to recognize and interact with other people because we can interpret their actions. Our brains enable all of this functionality, and they are unparalleled in their ability to process the world around us. Actions occur in three dimensions. As such, their perceived characteristics are affected by an observer's relative orientation and scale. Context also matters, as actions are highly variable based on the properties of the object performing the action as well as any objects that may be the target of the action. For example, chopping a soft vegetable like a tomato requires significantly less force than chopping a carrot.

The engineering of systems for human activity recognition in the field of computer vision has seen a dramatic growth over the last decade as evident by the number of publications and review articles [1], [2], [3]. Fueled by application needs in web-video search and retrieval, surveillance, health and wellness, human computer interfaces, and computer gaming, as well as advances in sensor technology, computing and algorithm development has resulted in impressive system performance in focused application domains. Crucial to the progress in the field was the development of standard databases in specific domains such as KTH (staged human actions in video) [4], UCF101 (human actions from videos in the wild) [5], HMDB51 (human motion recognition from video) [6], and VIRAT (activity recognition in surveillance video) [7]. Equally important are the open research community challenges and competitions such as LIRIS/ICPR2012 [8] and THUMOS [9].

Human actions occur in three-dimensional space and evolve over time. Most modern action recognition systems are based on visual data. Single RGB images capture a two-dimensional projection of the spatial arrangement of the human body in a scene. RGB video sequences capture the temporal evolution of those two-dimensional projections. Even more complete information can be gathered using RGB-Depth (RGB-D) videos that can provide the temporal evolution of a human body in three dimensions. Most of the state of the art systems are based on bag-of-features, local image descriptors derived from 2D images or 2D video volumes used in conjunction with a classifier, often a support vector machine [1]. The latter approaches, which employ low level features/representations, are simple and yield good results, but have the drawback in that they do not include prior knowledge about the spatial-temporal structure of the human body. Additionally, the visual/action words are not necessarily discriminative in the human action space. Thus current research in the field is moving towards more structured approaches using mid-level representations that are capable of capturing the complexity of real world applications. It is worth noting that the now popular deep network structures [10] can be viewed as creating discriminant mid-level representations [1].

In this paper we present a multimodal bio-inspired approach to action recognition inspired by the sonar systems of bats. Bats, which are the only mammals that can fly, have developed a sophisticated active sonar system that, coupled with their visual system [11], enable them to form structured representations of the complex environments that they reside in [12]. The horseshoe bat, unlike most bat species, has a constant frequency (CF) vocalization that allows it to detect and classify insects in cluttered environments [13]. More recently, perceptual experiments on the horseshoe bat suggest that the animals form a structured representation of their prey that relates to the physics of the prey's wing fluttering. By

T. S. Murray, D. R. Mendat, K. A. Sanni, P. O. Pouliquen and A. G. Andreou are with the Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, 21218.
E-mail: {thomasmurray, dmendat4, ksanni1, philippe, andreou}@jhu.edu,
Manuscript received March ??, 2016; revised XXX ??, 2016.

discerning the size of the prey, the bats are able to make intelligent decisions concerning prey selection by trading off the energy cost of flying to catch the prey with the benefits from the metabolic content of the prey [14]. By incorporating information from diverse sensory systems, biological systems are able to reliably identify the objects and actions they encounter.

Our approach to action recognition depicted in Figure 1 builds on a hidden Markov model (HMM) framework. Statistical models, trained simultaneously on both the micro-Doppler modulations induced by human actions and symbolic representations of skeletal poses, are developed. This enables the model to learn relations between the low dimensional, but rich, temporal structure of the micro-Doppler modulations and the high-dimensional pose sequences of human action from 3D video. During runtime, the model then relies purely on the active acoustic data to infer the human action. This approach utilizes a simple graphical model to capture the temporal sequences of skeletal poses and acoustic modulations and allows for the use of efficient inference algorithms. In Section II we describe the Doppler and micro-Doppler effects, followed by Section III where we outline our approach and experimental setup. Section IV describes the statistical action recognition model, followed by the results in Section V and discussion in Section VI.

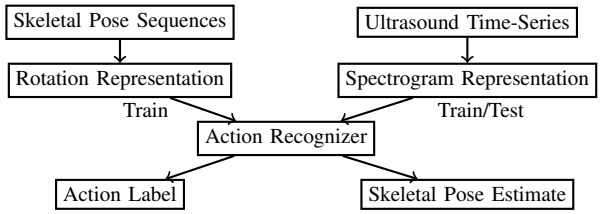

Fig. 1. Schematic of the proposed data flow in the action recognizer.

II. The Doppler and micro-Doppler Effects

In 1842, Christian Doppler postulated that the the frequency of waves emanating from a moving object relative to a stationary or moving observer would appear to be frequency shifted, a principle later named the Doppler effect [15]. While Doppler originally envisioned this principle being applied to electromagnetic waves (light), the first experimental observation of this phenomenon was done with acoustic waves by Buys Ballot [16] three years later. If the object itself contains moving parts, each part contributes its own Doppler shift proportional to the object's radial velocity component with respect to the receiver. All of the scattered waves are additive, and the resulting modulation is a superposition of the individual components known as the *micro-Doppler* effect [17]. The acoustic micro-Doppler effect was independently reported in 2007 by Zhang et. al. [18] and Kalgaonkar et. al. [19].

Assuming that there are N moving point masses in a scene where a pure tone with frequency f_c is transmitted, then the scattered signal seen by the receiver is

$$s_{\text{receiver}}(t) = \sum_{i=1}^{N} A_i(t) \cdot \sin(2\pi f_c t + 2\pi f_i t + \phi_i(t)). \quad (1)$$

Each point mass scatters the pure tone and modulates the frequency by $f_i = 2\frac{v_i}{c_s} f_c$, where v_i is the radial component of the velocity and c_s is the speed of sound. The amplitude of each component, A_i, depends on the scattering surface and the range of the point scatterer. There is also a phase shift $\phi_i(t)$ that depends on the range of the point mass. For ultrasound systems, the scattered wavelengths are typically on the order of 10mm, which allows relatively fine grained objects on the order of a couple millimeters to scatter the sound and produce modulations. Unfortunately, the short wavelength also means that the phase shift is not useful for extracting range information because it aliases after traveling a single wavelength. In comparison, micro-wave systems transmit wavelengths that are in the range of several centimeters. This means that acoustic systems are capable of resolving motion from smaller objects.

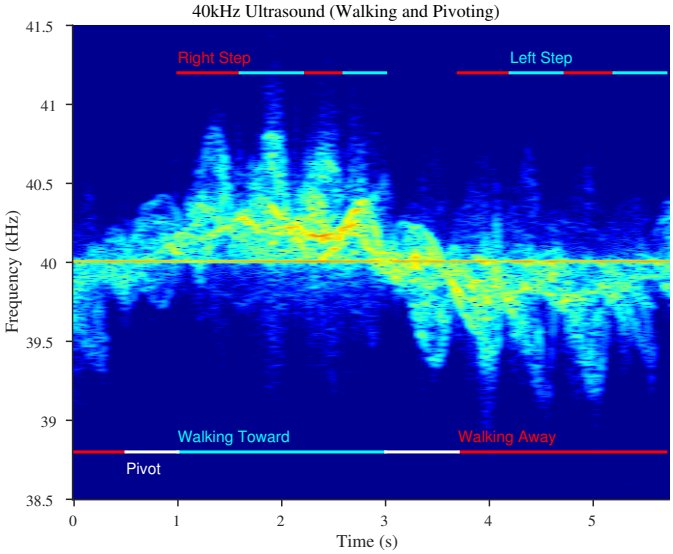

Fig. 2. Annotated spectrogram representation of Doppler modulations for a human walking toward an ultrasound sensor, pivoting, and walking back away from it.

The frequency spectrum of acoustic or electromagnetic waves scattered from a walking person is a complex time-frequency representation of human gait. In the case of a walking person, the torso, each arm section, and each leg section are all moving, and these individual movements each span a continuous range of velocities ranging from the slowest part (typically the proximal end) to the fastest (usually the distal end). The Doppler signature for such a complex object has infinite time-dependent frequency shifted components corresponding to the velocity ranges of the torso and individual limbs as a function of time. The time-domain micro-Doppler signal exhibits a complex structure that evolves in time. Therefore, a time-frequency spectrogram representation, consisting of a sequence of windowed spectral representations, namely the short-time Fourier transform (STFT), is more convenient for analyzing the changing spectrum of micro-Doppler signatures over time.

The electromagnetic micro-Doppler effect is exploited in radar applications [20] and in gait recognition experiments [21]. The analysis of electromagnetic signatures from humans in forest environments was recently reported [22].

The simplicity and cost-effectiveness of the sonar system in conjunction with its advantage in spatial resolution, which is millimeter for sound waves compared to centimeter for electromagnetic waves, has lead to the exploration of its use in different applications ranging from human identification and gender recognition [23],[24],[25],[26], [27], speaker identification [28], gesture recognition [29], transport mode [30], activity and behaviour classification [31], [32]. At this point, most of the efforts using sonar micro-Doppler are essentially pilot studies. This is partly because there have been no datasets comparable to the standard datasets in the vision community, which facilitate algorithm exploration in a systematic way.

III. Action Recognition Using the Micro-Doppler Effect

While the Doppler effect is very specific to sensing motion, there are still many challenges associated with exploiting it to sense and identify actions. At a fundamental level, real actions are sequences of motion that evolve in time and three-dimensional space. However, the micro-Doppler modulations recorded by a single active sonar sensor are one-dimensional time-series. The modulations of a pure tone used to sense a complex moving scene do not capture much in the way of range or spatial information. Over a given time window, the frequency modulations provide a histogram of the velocities present in the scene. Fortunately, due to the physical limitations of most multi-component objects, such as the human body, the realizable set of actions is heavily constrained. In the case of a human, the scattering components are linked rigid bodies, which constrain the space of human action and induce distinctive temporal structure in the modulations across a sequence of consecutive windows. This is a much more structured situation than the arbitrary sum of point masses expressed in Equation 1.

Another challenge is that many moving objects have symmetry in their motion. For example, a pendulum may swing from side to side and a human body may move its right or left arm. Distinguishing between these actions can be very challenging for a single sonar sensor, located at the line of symmetry, due to paucity of spatial information in the micro-Doppler modulations. One way to overcome this limitation is to use multiple sensor units arranged so that no single line of symmetry is common to all the sensors. In this section, we describe a new dataset of active acoustic and RGB-D recordings of human actions. The data was collected with a data acquisition system designed to integrate multiple acoustic sensors with very accurate temporal resolution. Leveraging this system allows for synchronized data collection with multiple sonar units that will help alleviate ambiguities due to spatial symmetry.

Although the space of possible human motions is quite large, there are a lot of constraints placed on actions by the physical limitations and structure of the human body. In theory, a model that captures the precise physical constraints of human joints and dimensions could be used to bias the decisions of an action recognizer that operates on impoverished acoustic signals. This approach leverages prior knowledge about the task and models the physics of the environment.

Additionally, the physics behind the Doppler effect are well understood. By incorporating prior knowledge about the interactions between the sensor and the environment, models can be developed that account for the interaction between the environment and the acoustics to extract as much information as possible from the data recorded by a given sensor. Models can also take advantage of the geometry of the sensor array in the environment to combine information from multiple sensors.

The Johns Hopkins University multimodal action (JHUMMA) dataset [33] is used in this study. Three ultrasound sensors [34] and a Kinect RGB-D sensor [35], [36] were used to record joint multimodal data of ten unique actors performing a set of actions. The dataset was created in an auditorium because it is a large open space and there are curtains on the stage where the data was collected. These features both reduce the number and strength of uninteresting reflections of the ultrasound carriers off static objects.

Figure 3 illustrates the configuration of the various sensors used for the data collection. The bounding box, which corresponds to the area where the Kinect sensor reliably tracks a human, was marked on the auditorium stage to guide the actors. All actions were confined to this space and the orientation of the actions and sensors is referenced to a virtual "North", which was defined as the orientation of an actor facing the Kinect sensor. The Kinect sensor was placed directly on top of the 40kHz ultrasound sensor (US40). The 25kHz ultrasound sensor (US25) was placed to the east and the 33kHz ultrasound sensor (US33) was placed to the west.

Figure 4 shows snapshots of the data recorded in the JHUMMA dataset during a single trial of each action. The first image associated with each action was captured by the Kinect sensor's color imager and the two-dimensional skeleton track has been superimposed on top of the image.

For human action recognition applications, it is desirable to develop algorithms capable of recognizing a particular action regardless of where it occurs in the global coordinate system. When training these algorithms it is advantageous to consider the hip-center as the origin for the skeleton at each frame. By referencing all of the other joints in a given frame to the position of the hip-center, the skeletal pose can be captured independently from the skeleton's global position. This skeletal pose representation provides translation invariance in the global coordinate system, which can greatly simplify the problem of recognizing a particular pose regardless of where a human is relative to the Kinect sensor. Storing the global position of the hip-center maintains all the necessary information to reconstruct the recorded scene exactly.

Furthermore, it is desirable if a human action recognition algorithm can be trained on skeletal poses collected from multiple subjects. One problem with the cartesian coordinates produced by the Kinect is their dependence on the height and limb lengths of the individual person. A very tall person and a very short person can perform the same action and generate very different cartesian joint coordinates even once the pose is adjusted to account for translation of the hip-center. However, the angle of the limbs as two people perform the same action is often much more consistent, even when their limbs are

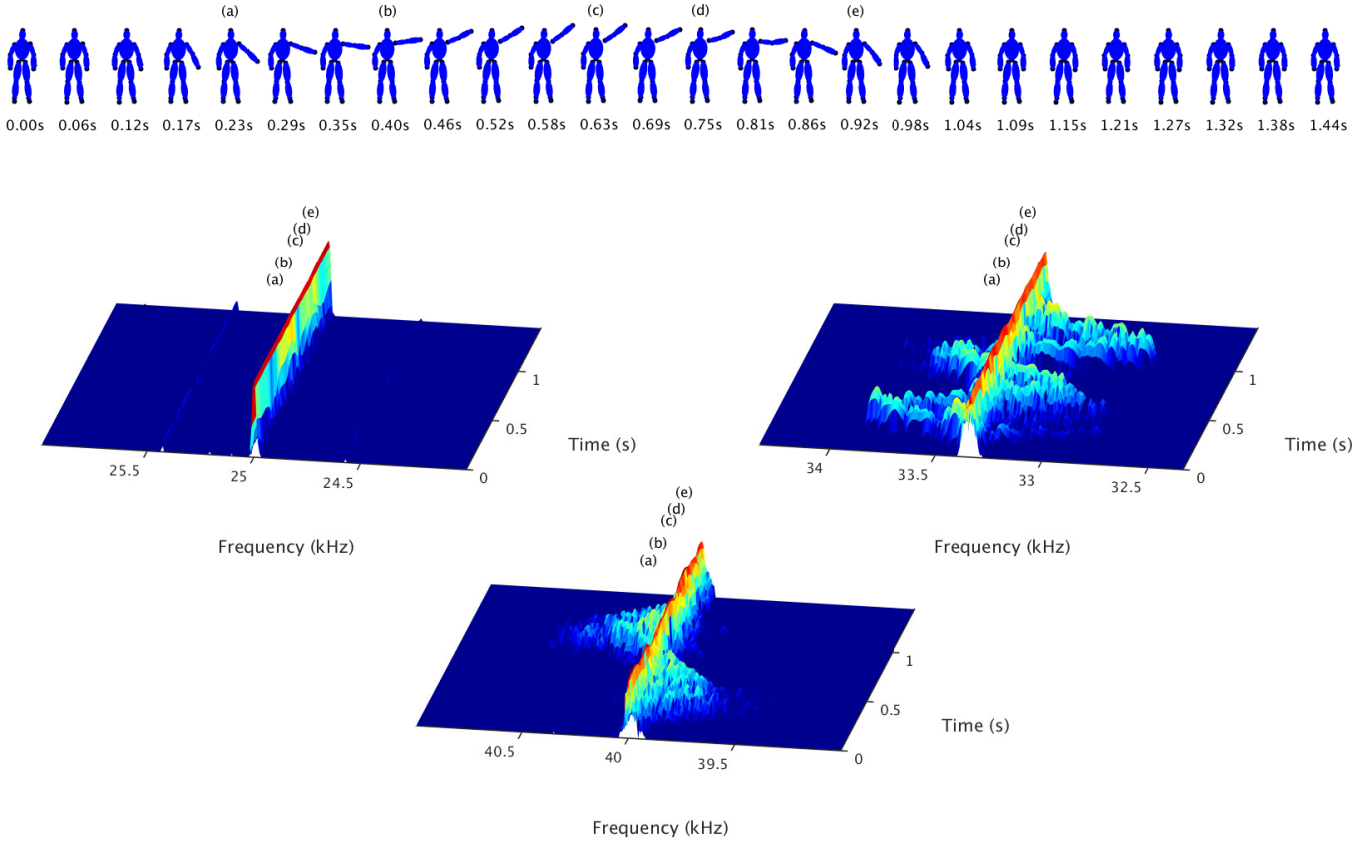

Fig. 3. Time evolution of action "Right-hand raise forward" and its representation in the continuous modulation spectra of the three micro-Doppler units. The absence of significant modulations in the 25kHz sensor (left spectrogram) due to occlusion from the body.

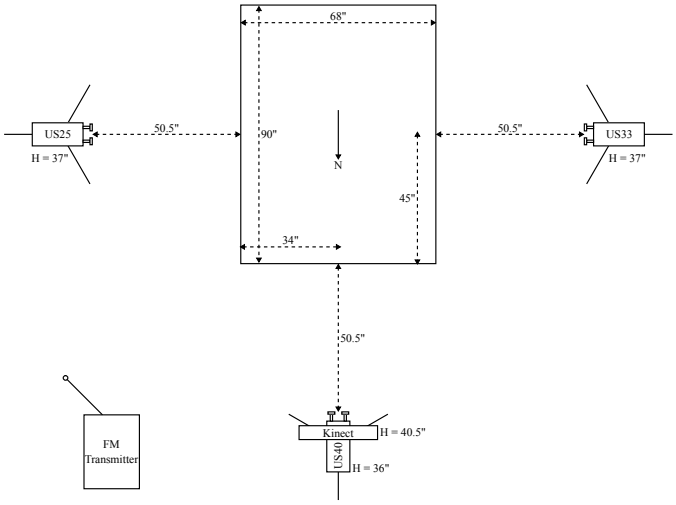

Fig. 4. Experimental setup for the JHUMMA data collection.

different lengths.

To leverage this invariance, the skeleton can be represented using the rotation of individual limbs instead of the cartesian coordinates of their constituent joints. The rotation representation is composed of two objects; an axis of rotation and an angle of rotation, θ. Figure 5 illustrates these components for a single limb. Each limb (blue line) is defined by two points, referred to as joint A and joint B. By convention, joint A is closer to the hip-center on the skeletal connection tree. The positive z-axis is used as a reference vector (red vector). The axis of rotation is the vector around which the limb must be rotated to match the reference. Due to the choice of reference, this axis is always constrained to the x-y plane.

IV. Action Recognition Model

Assuming that an appropriate dictionary of skeletal poses, $\mathcal{H}$, exists, the sequence of skeletal motion that results in human action can be approximated by $\mathbf{H} = h_0, ..., h_T$, where $h_t \in \mathcal{H}$. If we are also given an appropriate acoustic alphabet, $\mathcal{V}$ of acoustic spectrogram slices, then a spectrogram can then be described as $\mathbf{V} = v_1, ..., v_T$, where $v_t \in \mathcal{V}$. The methodology for generating the dictionary of skeletal poses and alphabet of acoustic modulations is developed later in Sections IV-E and IV-F. A set of action class labels, $\mathcal{C}$, enumerates the twenty-one actions in the JHUMMA dataset. Each sequence $\mathbf{H}$ is generated by an action, $a \in \mathcal{C}$, that modifies the parameters of the probability distributions accordingly.

The goal of the action recognizer is to estimate the most likely action that produced the visible sequence $\mathbf{V}$ of spectrogram slices. This can be expressed as

$$\begin{aligned} \hat{a} &= \arg\max_{a} \Big(\max_{\mathbf{H}} P_a(\mathbf{V}, \mathbf{H}) \Big) \\ &= \arg\max_{a} \Big(\max_{\mathbf{H}} P_a(\mathbf{V}|\mathbf{H}) P_a(\mathbf{H}) \Big), \end{aligned} \tag{2}$$

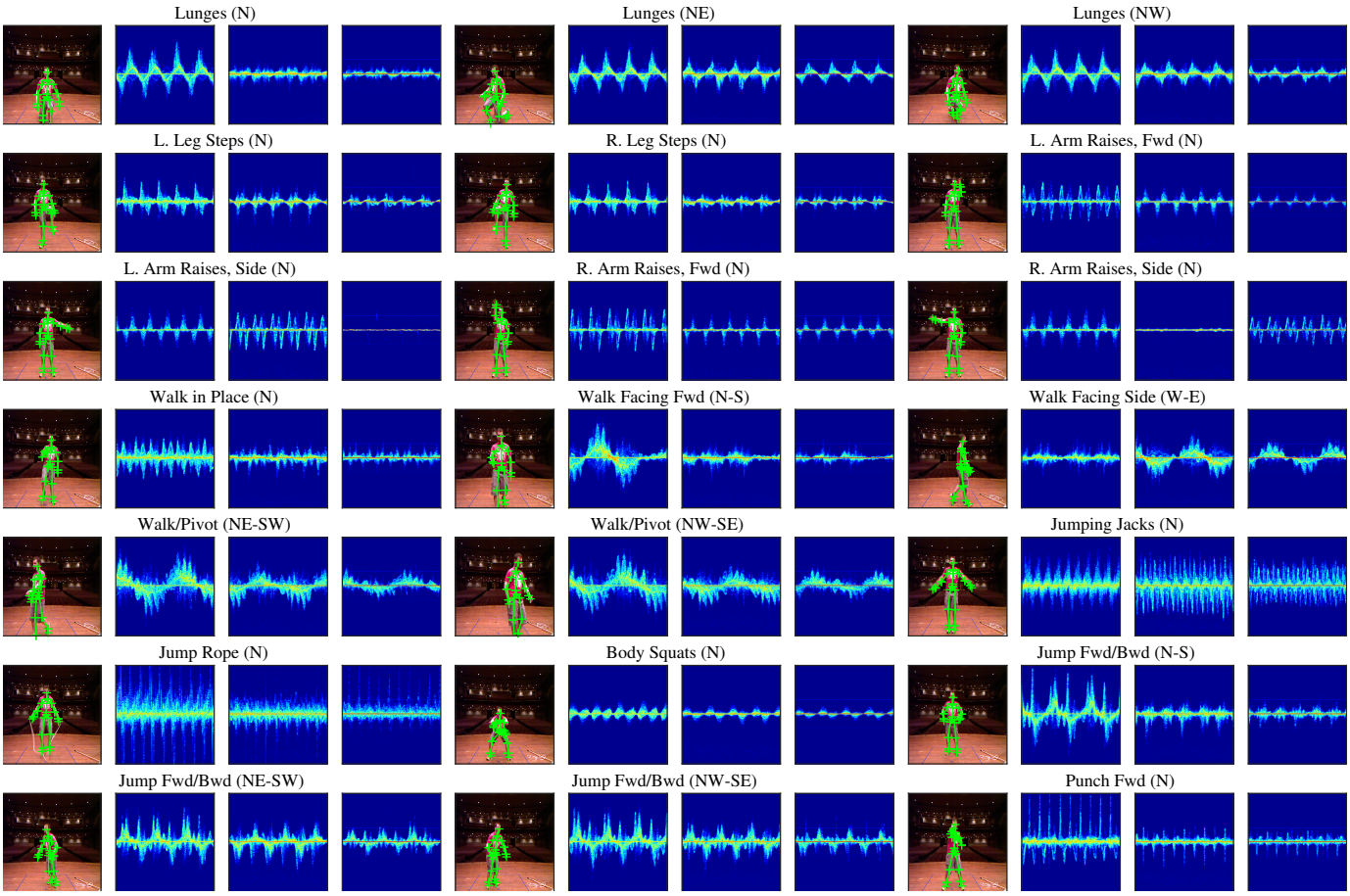

Fig. 5. Examples of the twenty-one actions contained in the JHUMMA dataset. An RGB color image, along with the 2D skeleton tracked by the Kinect sensor, as well as the acoustic modulations from each of the three ultrasound sensors, is shown for each action. The spectrogram in the second image was generated from ultrasound data recorded by the 40kHz sensor, the spectrogram in the third image was generated from ultrasound data recorded by the 33kHz sensor and the spectrogram in the fourth image was generated from ultrasound data recorded by the 25kHz sensor. The time window used to select the ultrasound data for each actions is the same for each sensor and the Kinect frames are all from within these windows. The window duration is fixed at just under nine seconds for each action. The displayed frequency content has a bandwidth of 2kHz centered on the respective carrier frequency of each ultrasound unit. The time and frequency markings are omitted for clarity.

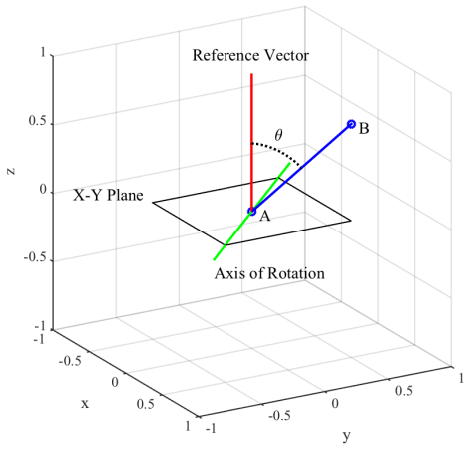

Fig. 6. Components of the rotation representation for a single limb.

where the joint distribution of a skeletal pose sequence and a spectrogram, $P_a(\mathbf{V}, \mathbf{H})$, is decomposed into a product of the skeletal pose sequence model, $P_a(\mathbf{H})$, and the active acoustic model, $P_a(\mathbf{V}|\mathbf{H})$, for a particular action class a.

A hidden Markov model (HMM) can be used to model a single pair of visible and hidden sequences. In order to leverage this model for recognizing actions, a set of HMM parameters are each trained separately on the portions of the training data that contain examples of a single action, a. When a new test sequence of acoustic spectrogram slices, $\mathbf{V}$, is observed, each of the individual action HMMs are used to predict the most likely sequence, $\mathbf{H}$, of unobserved skeletal poses. Computing the likelihoods of the sequences produced using each set of action parameters allows the models to predict the most likely action a by choosing the model that produces the most likely sequence. An HMM is an extension of a Markov chain where the random variables in the Markov sequence are considered hidden, or latent, and not observed. Instead, an additional visible random variable is observed at each step in the sequence [37]. The visible random variable is conditionally independent of all the other hidden and visible random variables given the hidden variable at that step. Figure 6 depicts the basic structure of the HMM used to

represent the active acoustic action model. HMMs have been extensively used in speech recognition [38], but their ability to capture the dynamics of the actions have also made them attractive for action recognition [39]. There are actually three independent sets of ultrasound observations in the JHUMMA dataset. In order to investigate the effects of using active acoustics from different orientations, three separate sets of HMMs are developed, one for each ultrasound sensor. While they can all share the same skeletal pose state space built upon the Kinect sensor data, their observation state spaces are all unique, requiring independent sets of parameters.

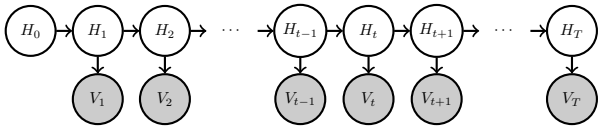

Fig. 7. Structure of the HMM used to capture sequences of Doppler modulations given a hidden sequence of skeletal poses.

In addition to the hidden sequence encoded by the variables $\mathbf{H} = h_0, ..., h_T$ and the visible sequence encoded by the variables $\mathbf{V} = v_1, ..., v_T$, the HMM also has a start state H_0 that is used to encode a set of prior probabilities indicating the likelihood that a chain starts in each state.

This HMM encodes the structure of the sub-motion and spectrogram slice sequences for a specific action a in the action recognition model if the sub-motions are assumed to adhere to the Markov property. Under this condition, the joint probability of the HMM can be decomposed as,

$$\begin{aligned} P_a(\mathbf{H}_a, \mathbf{V}) &= P_a(\mathbf{V}|\mathbf{H}) \cdot P_a(\mathbf{H}) \\ &= \prod_{t=1}^{T} P_a(V_t|H_t) \cdot P_a(H_0) \prod_{t=1}^{T} P_a(H_t|H_{t-1}) \ . \end{aligned} \tag{3}$$

The factorization of the joint distribution is also captured by the conditional independence assumptions encoded in the graphical structure of the HMM. The HMM is a generative probabilistic model of the active acoustic modulations and models the full joint probability distribution instead of just the discriminative class conditional distribution $P_a(\mathbf{H}|\mathbf{V})$.

The HMM parameters for action a are $\boldsymbol{\theta}_a = (\boldsymbol{\pi}_a, \mathbf{A}_a, \mathbf{B}_a)$, where $\boldsymbol{\pi}_a$ is the vector of hidden state priors, $\mathbf{A}_a$ is the matrix of transition probabilities between the hidden skeletal pose states and $\mathbf{B}_a$ is the matrix of emission probabilities of spectrogram slices from each of the hidden states. There are $|\mathcal{H}|$ hidden skeletal pose states and $|\mathcal{V}|$ visible spectrogram slice states. If $i \in \{1, ..., |\mathcal{H}|\}$ indexes into the the set of possible hidden states, then the elements of the state prior vector are,

$$\pi_a(i) = P_a(H_0 = i). \tag{4}$$

If $i, j \in \{1, ..., |\mathcal{H}|\}$ both index into the set of possible hidden states, then the elements of the transition matrix are,

$$A_a(i, j) = P_a(H_t = j|H_{t-1} = i). \tag{5}$$

If $i \in \{1, ..., |\mathcal{H}|\}$ indexes into the the set of possible hidden states and $k \in \{1, ..., |\mathcal{V}|\}$, then the elements of the emission matrix are,

$$B_a(i, k) = P_a(V_t = k|H_t = i). \tag{6}$$

A. Training the HMM Parameters

The JHUMMA dataset contains joint examples of both the hidden skeletal pose sequences and the visible spectrogram slices that can be used train the parameters for each class of actions. Under this fully supervised setting, the parameters for the HMM can be learned via closed-form maximum likelihood estimates (MLE) [40], [41], [42], [38]. To derive the MLE estimates, consider the joint probability of a training example, $(\mathbf{V} = \mathbf{v}, \mathbf{H} = \mathbf{h})$, where both the hidden and visible variables are known. Using Equation 3 gives the probability of the training example,

$$\begin{aligned} P_a(\mathbf{H} = \mathbf{h}, \mathbf{V} = \mathbf{v}) &= \prod_{t=1}^{T} B_a(H_t, V_t) \qquad (7) \\ &\times \prod_{t=1}^{T} A_a(H_{t-1}, H_t) \times \pi_a(H_0) \\ &= \prod_{t=1}^{T} \prod_{i=1}^{|\mathcal{H}|} \prod_{k=1}^{|\mathcal{V}|} B_a(i, k)^{\mathbb{I}(H_t=i, V_t=k)} \\ &\times \prod_{t=1}^{T} \prod_{i=1}^{|\mathcal{H}|} \prod_{j=1}^{|\mathcal{H}|} A_a(i, j)^{\mathbb{I}(H_{t-1}=i, H_t=j)} \\ &\times \prod_{i=1}^{|\mathcal{H}|} \pi_a(i)^{\mathbb{I}(H_0=i)}. \qquad (8) \end{aligned}$$

The parameters $\boldsymbol{\theta}_a$ have been substituted for the appropriate probability distributions and the indicator function $\mathbb{I}$ is used to specify the number of times each probability term occurs. The probability of L independent training sequences is simply $\prod_{l=1}^{L} P_a(\mathbf{H} = \mathbf{h}_l, \mathbf{V} = \mathbf{v}_l)$. sequences of training examples. Taking the log of this distribution yields,

$$\begin{aligned} \sum_{l=1}^{L} \log P_a(\mathbf{H} = \mathbf{h}_l, \mathbf{V} = \mathbf{v}_l) &= \sum_{i=1}^{|\mathcal{H}|} \sum_{k=1}^{|\mathcal{V}|} N_{ik} \log B_a(i, k) \\ &+ \sum_{i=1}^{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} N_{ij} \log A_a(i, j) \\ &+ \sum_{i=1}^{|\mathcal{H}|} N_i \log \pi_a(i). \end{aligned} \tag{9}$$

Here the emission counts across the training set are defined as,

$$N_{ik} = \sum_{l=1}^{L} \sum_{t=1}^{T} \mathbb{I}(H_{l,t} = i, V_{l,t} = k). \tag{10}$$

The transition counts across the training data are defined as,

$$N_{ij} = \sum_{l=1}^{L} \sum_{t=1}^{T} \mathbb{I}(H_{l,t} = j, H_{l,t-1} = i). \tag{11}$$

The prior counts across the training data are defined as,

$$N_i = \sum_{l=1}^{L} \mathbb{I}(H_{l,0} = i). \tag{12}$$

It is necessary to add additional constraints via Lagrange's multiplier. Essentially, the fact that the parameters are also

proper probability distributions, and therefore sum to unity, must be enforced. That is, $\sum_{i=1}^{|\mathcal{H}|} \pi_a(i) = 1$, $\sum_{j=1}^{|\mathcal{H}|} A(i,j) = 1$ and $\sum_{k=1}^{|\mathcal{V}|} B(i,k) = 1$. To find the MLE estimates for the various parameters, first add the appropriate constraint to the log-likelihood in Equation 9. Let λ be the Lagrange multiplier coefficient. Then take the partial derivatives of the constrained log-likelihood with respect to both the parameter of interest and λ. This results in two equations and two unknowns. For more details on using the Lagrangian to find the MLEs of the parameters, see Chapter 3 in Murphy [43]. Solving the system of equations for the state prior probabilities yields,

$$\hat{\pi}_a(i) = \frac{N_i}{\sum_{i'=1}^{|\mathcal{H}|} N_{i'}}. \tag{13}$$

Solving for the transition probabilities yields,

$$\hat{A}_a(i,j) = \frac{N_{ij}}{\sum_{j'=1}^{|\mathcal{H}|} N_{ij'}}. \tag{14}$$

Solving for the emission probabilities yields,

$$\hat{B}_a(i,k) = \frac{N_{ik}}{\sum_{k'=1}^{|\mathcal{V}|} N_{ik'}}. \tag{15}$$

Under the supervised training paradigm, finding the MLE estimates for the HMM parameters essentially boils down to counting the number of times the relevant event occurred in the training data and normalizing the results into proper distributions. In order to train one set of HMM parameters for each action $a \in \mathcal{C}$, the training data is split according to the action that generated it and the parameters for each action are trained solely on the associated training data.

Many of the possible hidden state transitions and visible observation combinations were never observed in the training sets. To alleviate this, add-one smoothing was applied to the MLE estimates. This technique amounts to adding one phantom count to each element prior to normalization.

B. Finding the Most Likely Hidden Sequence in a Hidden Markov Model

Given the trained parameters for an HMM and a test sequence of observations, the Viterbi algorithm [41], [38], [44] can be used to find the most likely sequence of hidden states. The Viterbi algorithm is a dynamic programming technique to efficiently compute the maximum a posteriori (MAP) probability estimate of the most likely sequence in a chain-structured graphical model, such as the HMM.

The Viterbi algorithm is composed of a forward pass through all possible sequences of states where the likelihood of ending up in state $j \in \{1, ..., |\mathcal{H}|\}$ at time $t \in \{1, ..., T\}$ is computed for each state. Given an observed sequence $V_1 = k_1, ..., V_T = k_T$, the likelihood $\delta_t(j)$ of a state j, at each time step t, can be computed based on the likelihoods of the states at the previous time step $t-1$, the transition probabilities between the states and the probability that each state emits the current observed symbol k_t,

$$\delta_t(j) = \max_{i=2,...,|\mathcal{H}|} \delta_{t-1}(i) A(i,j) B(j,k_t). \tag{16}$$

The forward pass can be initialized using the prior probability of each state such that,

$$\delta_1(j) = \max \ \pi(i) A(i,j) B(j,k_1). \tag{17}$$

In addition to tracking the likelihood of each state, the previous state that gave rise to the likelihood is also tracked.

$$\alpha_t(j) = \underset{i=1,...,|\mathcal{H}|}{\arg\max} \ \delta_{t-1}(i) A(i,j) B(j,k_t). \tag{18}$$

The Viterbi algorithm for an HMM terminates once the final time step T is reached. At this point the sequence of most likely states can be traced backwards through time. Beginning at time step T, the most likely state is

$$h_T^* = \underset{i=1,...,|\mathcal{H}|}{\arg\max} \ \delta_T(i), \tag{19}$$

and the sequence is unrolled using the previous states that were tracked. Thus,

$$h_t^* = \alpha_{t+1}(h_{t+1}^*), \tag{20}$$

where $t < T$.

C. Splitting the JHUMMA Dataset into Examples and Batches

The JHUMMA dataset provides a perfect foundation for building HMMs that jointly model sequences of skeletal poses and sequences of Doppler-modulations and to evaluate their ability to classify actions sequences. There are 21 distinct types of actions captured in the JHUMMA dataset and the performance of the HMM model is evaluated on the task of classifying these action categories.

The JHUMMA dataset contains a sequence of spectrogram slices for each of the three ultrasound sensors and a sequence of skeletal frames for each of the actions performed during each of the thirteen trials. Unfortunately, each of these coarsely labeled sequences contains multiple repetitions of the same action. Nominally each sequence contains ten repetitions, although there are several instances where the actor lost track of the count. In order to generate test sequences suitable for training and testing HMMs, each sequence was split into ten examples of equal numbers of consecutive frames. Any remaining frames were appended to the last example so that temporal cohesion is maintained.

Five batches of training and testing data were set up for cross-validation. For each action/trial pair, two of the ten data sequences were randomly selected as test examples, while the remaining eight were selected as training examples. The random permutations were constructed such that each of the ten examples serves as a test sequence in exactly one of the five batches. One of the actors accidentally skipped three actions, so there are precisely $2,160$ training examples and 540 test examples in each batch.

D. Learning Cluster Prototypes

The K-means algorithm is a common method for performing vector quantization [45], [43], a technique for modeling probability densities based on the location of prototype vectors. The idea behind K-means was first proposed by Steinhaus as

least squares quantization in pulse-code modulation (PCM) and the standard algorithm used to implement the technique was first published by Lloyd [46] [47] with efficient large scale applications of the algorithm advanced by Coates [48] and Kanungo [49]. In Sections IV-E and IV-F the details of using K-means to learn prototype clusters for both the skeletal poses and spectrogram slices are described. Here the basic algorithm is developed for performing unsupervised clustering.

Let $\mathbf{X} = \mathbf{x}_1, ..., \mathbf{x}_N$ be a set of unlabeled training data, where each $\mathbf{x}_i \in \mathbb{R}^D$. Define a set of $j = \{1, ..., K\}$ cluster prototypes $\boldsymbol{\mu}_j \in \mathbb{R}^D$. The cluster prototypes are initialized using the K-means++ algorithm [50], which randomly selects one of the data points from $\mathbf{X}$ to be the first cluster prototype and selects subsequent points, one at a time, from $\mathbf{X}$ to be initial cluster prototypes with probability inversely proportional to their distance from the nearest existing selected prototype [49].

Once all K of the clusters have been initialized, the training data points are all assigned to the nearest cluster. In this work, the distance between any data point in the training set and any cluster mean is given by

$$d(\mathbf{x}_i, \boldsymbol{\mu}_j) = \sqrt{\sum_{d=1}^{D} (x_d - \mu_d)^2}, \tag{21}$$

the Euclidean distance in D-dimensional space. Once the cluster assignment is complete, the cluster prototypes are updated by computing the mean value of all the data points in the cluster assignment. Then, using these new cluster prototypes, the procedure is repeated. The stopping criterion is generally when no data points change clusters in successive iterations.

In this work, the K-means algorithm was performed four times with different random data points used for the K-means++ cluster initialization. The decision to use four starting points was based on the number of available independent CPU cores. The number of iterations was also capped at 750, although the cluster prototypes converged before that in all cases.

E. Skeletal Pose State Space Model

One limitation of the HMM is that it is built on a finite state space. Unfortunately, the skeletal poses derived from the Kinect data are more accurately represented in continuous space. In order to generate a finite latent space of skeletal poses suitable for training the HMMs, we employ the K-means algorithm to discover unsupervised clusters suitable for quantizing the vector space of skeletal poses.

Ideally, the model for the hidden state variables would capture the skeletal pose precisely at a given instant in time. However, one limitation of the HMM is that the state space is finite, so there must be a finite number of hidden states. The approach taken in this work is to find a set of skeletal poses that suitably approximate the space of skeletons recorded by the Kinect sensor in the JHUMMA dataset. This was accomplished through unsupervised clustering, using the K-means algorithm described in Section IV-D, to find cluster prototypes given all of the skeletal frames in the training data of a given batch. The process was then repeated separately for the training data in each of the cross-validation datasets. The principle parameter involved with this method is the degree to which the training skeletons are quantized, which is the number of skeletal clusters, K. The hidden state variables H_t take on values $h_t \in \{1, ..., K\}$, which index the set of skeletal pose clusters.

The skeletal poses from the Kinect were adapted in three ways to simplify the problem and facilitate clusterings that capture the most important information. The first adaptation was to remove the translation of the hip joint from the features included in the skeletal clusters. As discussed in Section III, this provides translation invariance, which is critical so that the pose clusters that are learned are applicable to any location in the dataset. It would be prohibitively expensive to produce and label a dataset extensive enough to support learning individual clusterings for different spatial areas.

The second adaptation was to remove the hand and feet joints from the skeletal poses. Studying the Kinect data in the JHUMMA dataset reveals that the hands and feet tend to be the noisiest joint estimates. The feet in particular tend to exhibit a significant amount of jitter from frame to frame. Removing these joints prevents the learned skeletal clusters from spending any modeling power accounting for the jitter in these noisy joints. It also has the added benefit of reducing the dimensionality of the skeletal pose features, which is also the dimension of the cluster space. Removing the hands and feet left only 16 joints in the abbreviated skeleton structure.

The third adaptation was to use the rotation representation of the skeletal pose, described in Section III. This allows all of the training data, regardless of the actor, to be merged together. The skeletal poses are clustered in limb rotation space, which is more amenable to cross-training between actors than cartesian joint coordinates. The limb rotations are referenced to the vertical and only take on values in the range of 0 radians, which corresponds to straight up, to π radians, which corresponds to straight down. In this representation, the discontinuity between 0 radians and 2π radians is avoided, so the Euclidean distance remains a natural choice of metric. Applying all three of these adaptations resulted in each skeletal pose being represented by a 45-dimensional feature vector.

In order to explore the effect of different quantization levels in the pose space, the K-means clustering procedure was performed for various numbers of clusters on the first batch of data. Figure 7 shows the average joint error for each set of skeletal pose clusters. The error was calculated by computing the distance between each joint in each skeletal frame of the training data and the corresponding joint in the cluster mean that the training frame was associated with. For the purposes of computing the error, the rotation representation of the cluster mean was transformed back into cartesian coordinates. The error was summarized by averaging across each of the 16 joints in each of the $225,483$ training frames, which were pooled across all thirteen trials and 21 actions in the first batch of data.

The curve in Figure 7 illustrates a tradeoff between model complexity and accuracy. As the number of skeletal clusters increases, the clusters do a better job of approximating the

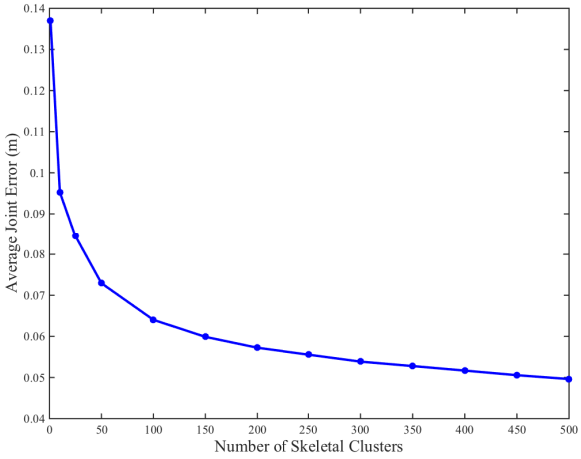

Fig. 8. Comparison of the error between each skeletal frame in the training data and the associated skeletal cluster for various numbers of clusters.

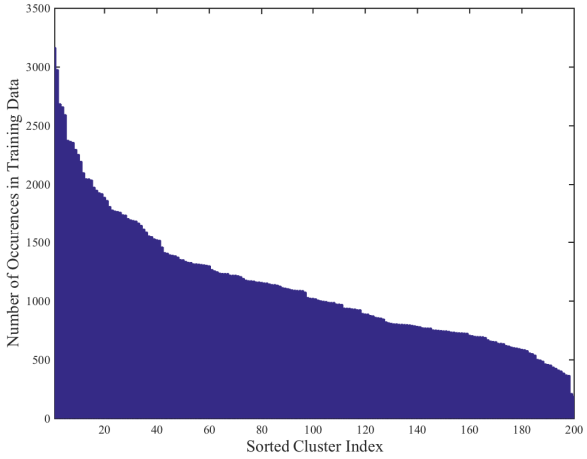

Fig. 10. Histogram illustrating the number of occurrences of each skeletal cluster. The cluster indices have been sorted by their frequency.

training data, so the error decreases. However, more clusters require more model parameters to be estimated. Unless otherwise specified, the data shown in the following sections was generated using 200 skeletal clusters, which errs on the side of accurately modeling the skeletal poses with a more complex model.

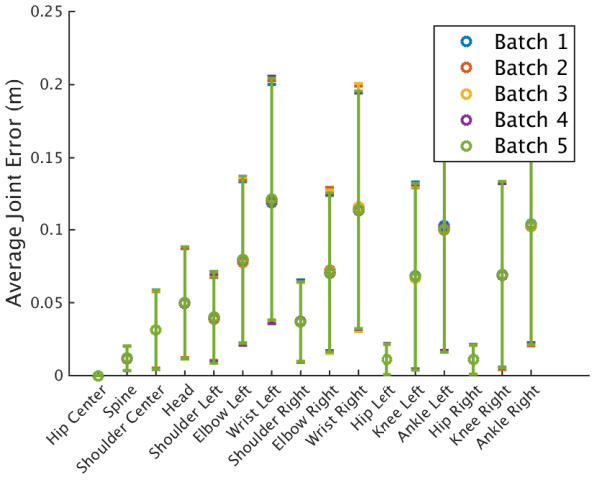

Fig. 9. Comparison of the error between each skeletal joint in the training data and the closest skeletal cluster for each data batch. The number of skeletal clusters was fixed at 200.

In order to confirm that the training data in each cross-validation batch produces similar quantization results, the error of each joint was investigated. The error was computed as the Euclidean distance from each joint in the training data relative to the corresponding joint in the associated skeletal cluster. Figure 8 shows the error for each of the 16 joints, averaged across all of the training examples in each of the five batches. The error bars in Figure 8 correspond to one standard deviation of the joint errors.

As mentioned earlier, the hip translation was removed from the representation, so all of the hip joints were fixed to the origin when the other joint errors were computed, which is why they appear to have zero error. It is also interesting to note that the wrist and ankle joints have significantly higher error and variance than the others. This makes sense because they tend to move more during actions. They are also more likely to be tracked erroneously by the Kinect. This result supports the decision to omit the hand and foot joints, which were even more unreliable.

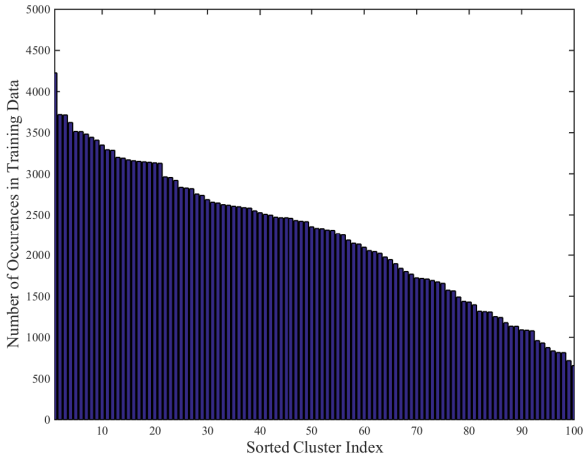

Fig. 11. Histogram illustrating the number of occurrences of each 40kHz ultrasound cluster. The cluster indices have been sorted by their frequency.

Figure 10 shows the frequency of each 40kHz ultrasound cluster extracted from the first batch of cross-validation data. The cluster frequencies appear reasonable. Some are certainly more frequent than others, but no cluster dominates.

One nice feature of the spectrogram slices is that they are relatively easy to display and interpret in two-dimensions. Figure 11 shows all of the cluster means for the 40kHz training data. These representative spectrogram slices are sorted according to their frequency in the training data. Actions are still more often composed of periods of little movement, with

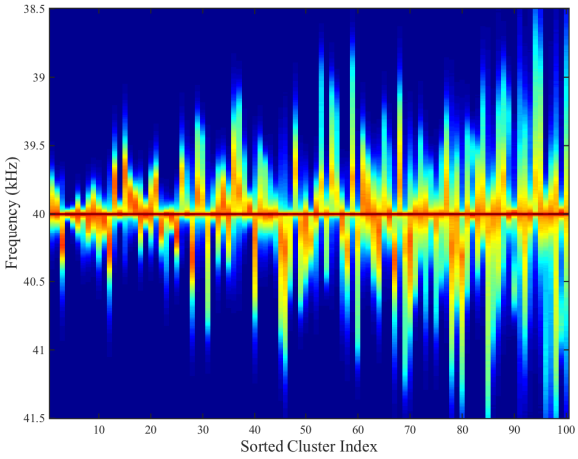

Fig. 12. The collection of 40kHz ultrasound representative spectrogram slice cluster means, ordered by their frequency in the training data.

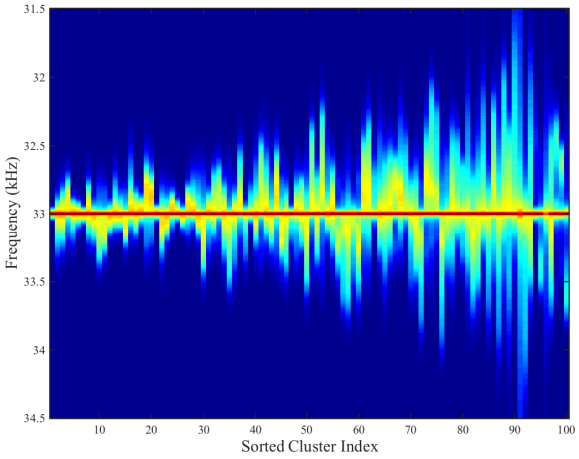

Fig. 14. The collection of 33kHz ultrasound representative spectrogram slice cluster means, ordered by their frequency in the training data.

large motions being relatively rare, which parallels the general trend of clusters with larger Doppler modulations being less frequent.

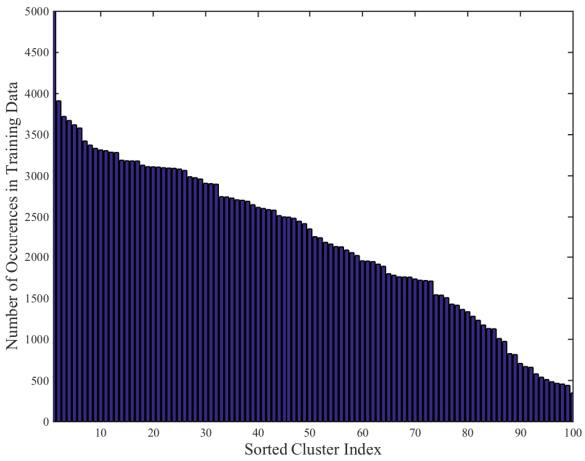

Fig. 13. Histogram illustrating the number of occurrences of each 33kHz ultrasound cluster. The cluster indices have been sorted by their frequency.

Figure 12 shows the frequency of each 33kHz ultrasound cluster extracted from the first batch of cross-validation data. The distribution is slightly more skewed than the one for the 40kHz data.

Figure 13 shows all of the cluster means for the 33kHz training data. The clusters are very similar in character to those culled from the 40kHz data. The modulations are smaller overall, but this is due to the lower carrier frequency and the fact that the sensor was positioned to the side of the majority of the actions in the JHUMMA dataset. The side sensors tended to observe smaller velocity components for the majority of actions. This is also supported by the histogram of the clusters, which indicates that the higher modulation clusters, indicative of more motion towards the side sensors, are less frequent compared to the histogram of the 40kHz sensor, which was positioned directly in front of most of the actions.

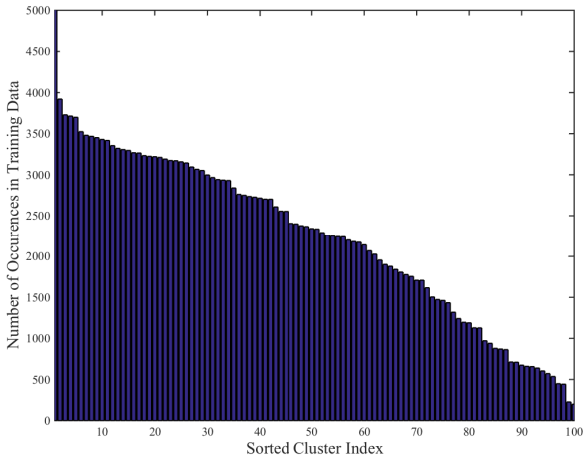

Fig. 15. Histogram illustrating the number of occurrences of each 25kHz ultrasound cluster. The cluster indices have been sorted by their frequency.

Figure 14 shows the frequency of each 25kHz ultrasound cluster extracted from the first batch of cross-validation data. Similarly to the 33kHz spectrogram slice clusters, the 25kHz spectrogram slice clusters also appear to have a more skewed distribution than the 40kHz spectrogram slice clusters. This is in line with the less variable nature of both the positioning of the sensor off to the side and the lower magnitude of the 25kHz modulations. Figure 15 shows all of the cluster means for the 25kHz training data.

Figure 9 shows the frequency of each skeletal pose cluster in the training data for the first cross-validation batch. The cluster indices are sorted according to their frequency. Although the frequency is not uniform, the balance between cluster frequencies appears reasonable. Some actions have relatively unique skeletal poses that are not exhibited often, while many

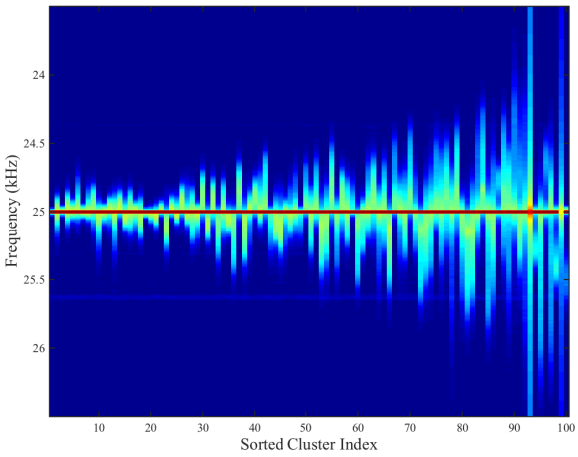

Fig. 16. The collection of 25kHz ultrasound spectrogram slice cluster means, ordered by their frequency in the training data.

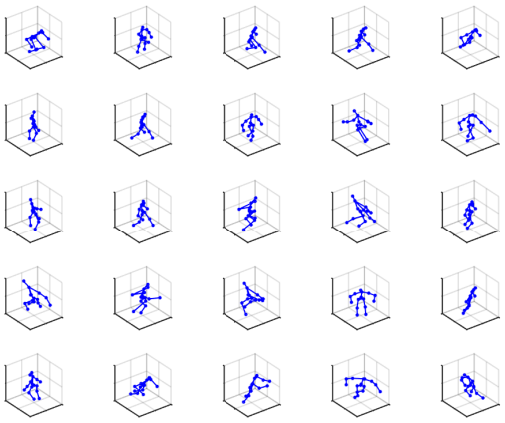

Fig. 17. A random sampling of 25 of the 200 skeletal clusters learned from the first batch of training data.

actions share similar skeletal poses that are clustered together and occur more frequently.

Figure 16 shows a random sampling of the skeletal pose clusters learned from the first batch of cross-validation data. These poses appear to be relatively diverse and interesting, indicating that the unsupervised clustering approach is at least reasonable.

F. Doppler Modulation Observation Model

While the hidden variables for each of the three HMM models can all utilize the same set of skeletal pose clusters, it is necessary to develop sets of spectrogram slice clusters that are tuned to each of the three ultrasound sensors individually because they each utilize a different carrier frequency.

An approach similar to the skeletal clustering was taken to quantize the spectrogram slices associated with each ultrasound sensor. The spectrogram slices from all of the training sequences were pooled together and the K-means algorithm was again used to choose a set of average clusters that were representative of the entire set. Figure 17 shows the average error for a spectrogram slice in the first batch of 40kHz ultrasound data over several values of K. The average cluster error was also computed using clusterings derived using both the Euclidean, or L_2, and L_1 distance metrics. Although the L_2 distance metric is not an obvious choice for comparing two spectrogram slices, empirical testing demonstrated very little difference between the character or performance of spectrogram clusters created using the L_2 distance metric versus the L_1 distance metric.

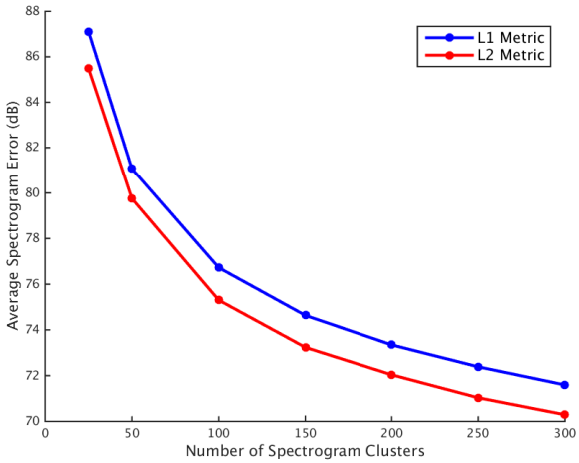

Fig. 18. Comparison of the average error between each 40kHz spectrogram slice in the first batch of training data and the nearest spectrogram cluster for increasing numbers of cluster prototypes, K. The clustering K-means clustering procedure was run using both the L_1 and L_2 distance metrics.

For clustering the ultrasound spectrogram slices, a value of $K = 100$ was used. As the clusterings do not appear to be particularly sensitive to the choice of distance metric, the spectrogram clusters used to generate the results presented here were created using the L_2 distance metric, which is consistent with the metric used to cluster the skeletal poses. This spectrogram clustering process was repeated separately for the data from each of the three ultrasound sensors and for each of the cross-validation datasets.

V. Human Action Recognition Results

Once an appropriate vocabulary of skeletal pose prototypes was constructed from the training data and the alphabets of spectrogram slice prototypes were learned separately for each of the ultrasound frequencies, the parameters for each of the actions classes and ultrasound sensors were computed using Equations 13, 14 and 15. To classify a novel test sequence from one of the ultrasound sensors, it was first translated into a sequence of spectrogram slice prototypes. This was done by choosing the prototype with the smallest Euclidean distance from each spectrogram slice in the test sequence. Once the test data was translated into spectrogram prototypes $\mathbf{v}$, the most likely sequence of hidden skeletal pose prototypes $\mathbf{h}_a^*$ was computed using the Viterbi algorithm, described in Section IV-B. This procedure was repeated for each set of

parameters $\boldsymbol{\theta}_a$. Note that only the parameters trained for that particular ultrasound frequency were considered, and the subscript on the most likely hidden sequence was used to indicate the action the set of HMM parameters used to produce it was trained on.

The log-likelihood of a hidden sequence $\mathbf{h}$ and an observed sequence $\mathbf{v}$, normalized for the number of time steps in the sequences, is

$$\begin{aligned} \mathcal{L}_a(\mathbf{h},\mathbf{v}) &= \log \pi_a(h_0) \\ &+ \sum_{t=1}^{T} \log A(h_{t-1},h_t) \\ &+ \sum_{t=1}^{T} \log B(v_t,h_t) - \log T. \end{aligned} \tag{22}$$

After computing the log-likelihood of the hidden sequence produced by each action model, the sequence was classified as the action that best modeled the sequence. That is,

$$\hat{a} = \arg\max_a \mathcal{L}_a(\mathbf{h},\mathbf{v}). \tag{23}$$

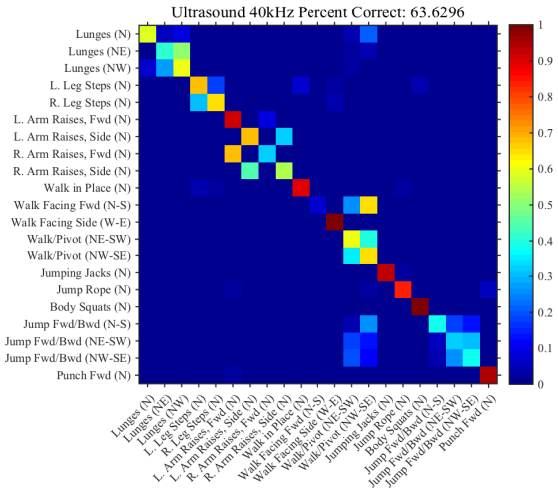

Fig. 19. Confusion matrix enumerating the action classification decisions resulting from the 40kHz ultrasound model.

Figure 18 shows the confusion matrix for the action classification task that results from HMMs trained on the 40kHz ultrasound data. Overall, the HMM model correctly classified 63.63% of the 2700 test examples in the JHUMMA dataset. There were twenty-one actions, so classifying the actions by chance would yield a classification rate of under 5%. These results were compiled using all five of the cross-validation batches.

The confusion matrix indicates that the model tends to make very specific types of errors. It has significant difficulty distinguishing left versus right orientation among the action classes that have a similar action but different orientation. This is evident by the blocks of misclassification errors that are formed around many of the actions that have multiple orientations. One such example is the classification of the left leg steps and the right leg steps. The classifier places almost all of the probability mass on one of those two actions, but there is a lot of error between them. Recall that the 40kHz ultrasound sensor was positioned to the north of the actor, which is roughly the line of symmetry for left versus right actions. With limited spatial information in the modulations, distinguishing between arm raises to one side or the other is difficult and results in significant classification errors. On the other hand, the 40kHz ultrasound HMM does a good job of predicting actions with unique orientations such as punching and jumping jacks. This indicates that the modulations themselves are reasonable informative patterns to use for classifying coarse-grained action sequences.

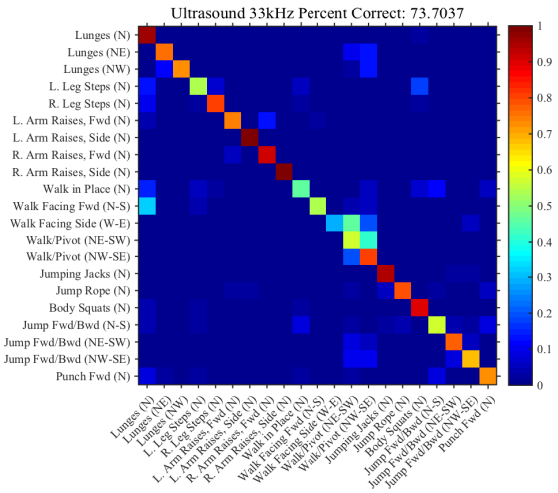

Fig. 20. Confusion matrix enumerating the action classification decisions resulting from the 33kHz ultrasound model.

Figure 19 shows the confusion matrix for the action classification task that results from HMMs trained on the 33kHz ultrasound data. Overall, the HMM model correctly classified 73.70% of the 2700 test examples in the JHUMMA dataset. Almost all of the actions with multiple orientations were symmetric with respect to the North to South axis of the JHUMMA setup. Therefore, it makes sense that the HMM trained on the micro-Doppler modulations recorded by the 33kHz ultrasound sensor, which was off to the West, made fewer errors than the 40kHz ultrasound sensor. In fact, the one set of actions that did have some orientations facing the 33kHz sensor, walking back and forth, exhibited the same block error patterns in the confusion matrix as are evident in the 40kHz ultrasound classifications.

Figure 20 shows the confusion matrix for the action classification task that results from HMMs trained on the 25kHz ultrasound data. Overall, the HMM model correctly classified 75.30% of the 2700 test examples in the JHUMMA dataset. The errors made by the HMM model trained on data from the 25kHz ultrasound sensor, which was positioned to the East, is qualitatively similar to the errors made by the 33kHz ultrasound sensor. This is reasonable as both sensors were on the same cardinal axis and, therefore, encountered the same ambiguities due to the orientation of the actions in the JHUMMA dataset.

Given that the position of the ultrasound sensor has a significant effect on the classification accuracy of the model trained on data recorded by it, a fourth model that is a fusion of all three individual ultrasound HMMs was created to investigate the benefits of combining multiple ultrasound

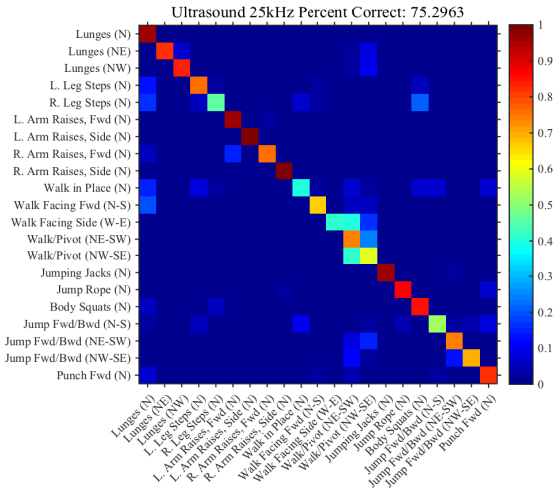

Fig. 21. Confusion matrix enumerating the action classification decisions resulting from the 25kHz ultrasound model.

sensors to disambiguate the orientations of the actions. The combined model was constructed as a product of the individual models by summing the log-likelihoods for each action that were produced by the individual ultrasound models prior to choosing the most likely action. A test sequence $\mathbf{v}$ is now classified based on,

$$\hat{a} = \arg\max_{a} \left(\mathcal{L}_a^{40}(\mathbf{h}, \mathbf{v}) + \mathcal{L}_a^{33}(\mathbf{h}, \mathbf{v}) + \mathcal{L}_a^{25}(\mathbf{h}, \mathbf{v}) \right). \quad (24)$$

Figure 21 shows the confusion matrix for the action classification task that results from combining each of the individual ultrasound HMM as a "product of experts" model. Overall, the HMM model correctly classified 88.56% of the 2700 test examples in the JHUMMA dataset. Combining the output of the individual HMM models gives a significant boost in classification performance and appears to be a reasonable approach to leveraging multiple ultrasound sensor units.

VI. Discussion

Table I gives a more detailed breakdown of the exact classification rates for each of the three individual ultrasound models as well as the product of experts model combining them all. Table II presents a comparison of classification performance for several different numbers of skeleton pose cluster prototypes. On the left side, the overall classification results for the action sequences are shown. On the right side, the pose classification rate for the hidden sequence of cluster prototypes predicted from the data of each ultrasound band are shown. The pose classification rate is computed by comparing the closest skeletal pose prototype, at each time step in a test sequence, to the skeletal pose prototype predicted by the HMM given the test sequence of ultrasound modulations.

In general, more skeletal pose prototypes result in a more expressive state space that is able to model the actual recorded skeletal poses more closely. However, this precision comes at the price of a significantly larger model that now has many more parameters to estimate from the same fixed pool of training data. This is a classic model selection tradeoff and the results in Table II illustrate this. The action classification rates generally increase with the number of skeletal pose prototypes. However, the pose classification rates increase with fewer skeletal pose prototypes. This is reasonable because fewer prototypes make estimating the closest one significantly easier. Overall, using 200 skeletal pose prototypes, the conclusion drawn from the tradeoff in Figure 7, seems to be a reasonable compromise between these two trends.

The work presented in this paper, employs fundamental mathematical models and tools that are also employed in the vision community aimed at capturing the dynamics and kinematics of human body. Statistical models such as HMM and CRFs [51],[52],[53],[54] as well as the work that employs dynamic probabilistic networks [55], [56], [57] incorporates more structure and prior knowledge in the recognition process through temporal, contextual and ordering constraints in the models. Also relevant is the work on linear [58] and non-linear dynamical systems approach [59] on tracked features and or optical flows are alternative methods aimed at recognizing activities that are concatenation of simpler actions. The more advanced models in the latter body of work in the computer vision community could be applied to the problem and sensor data in this paper to further improve the performance of the action recognition system.

TABLE I
Action Classification Performance on the JHUMMA Dataset.

Action Label	25 kHz	33 kHz	40 kHz	Com-bined	Test Ex-amples
Lunges (N)	95.38	95.38	59.23	100.00	130
Lunges (NE)	81.54	75.38	40.77	91.54	130
Lunges (NW)	83.85	73.08	60.77	87.69	130
L. Leg Steps (N)	75.83	53.33	67.50	87.50	120
R. Leg Steps (N)	46.67	80.83	64.17	93.33	120
L. Arm Raises, Fwd (N)	96.15	73.85	90.77	100.00	130
L. Arm Raises, Side (N)	99.23	100.00	68.46	100.00	130
R. Arm Raises, Fwd (N)	76.15	91.54	31.54	96.92	130
R. Arm Raises, Side (N)	100.00	99.23	54.62	100.00	130
Walk in Place (N)	39.23	45.38	89.23	85.38	130
Walk Facing Fwd (N-S)	66.92	53.85	7.69	69.23	130
Walk Facing Side (W-E)	40.77	29.23	98.46	89.23	130
Walk/Pivot (NE-SW)	74.62	57.69	60.77	64.62	130
Walk/Pivot (NW-SE)	58.46	80.77	65.38	78.46	130
Jumping Jacks (N)	95.38	93.85	93.08	98.46	130
Jump Rope (N)	86.15	78.46	83.85	84.62	130
Body Squats (N)	84.62	89.23	100.00	100.00	130
Jump Fwd/Bwd (N-S)	53.08	56.92	37.69	83.85	130
Jump Fwd/Bwd (NE-SW)	73.85	77.69	32.31	76.92	130
Jump Fwd/Bwd (NW-SE)	70.00	68.46	37.69	74.62	130
Punch Fwd (N)	81.67	72.50	95.00	98.33	120
Overall	75.30	73.70	63.63	88.56	2700

TABLE II
Action and Pose Classification Performance on the JHUMMA Dataset.

Number of Clusters	Action Classification			Pose Classification		
	25kHz	33kHz	40kHz	25kHz	33kHz	40kHz
100	73.56	72.89	60.63	25.36	25.53	20.91
150	74.07	74.00	64.00	23.34	23.25	19.80
200	75.30	73.70	63.63	22.12	21.80	17.97
300	75.11	74.63	62.67	19.30	19.39	15.82

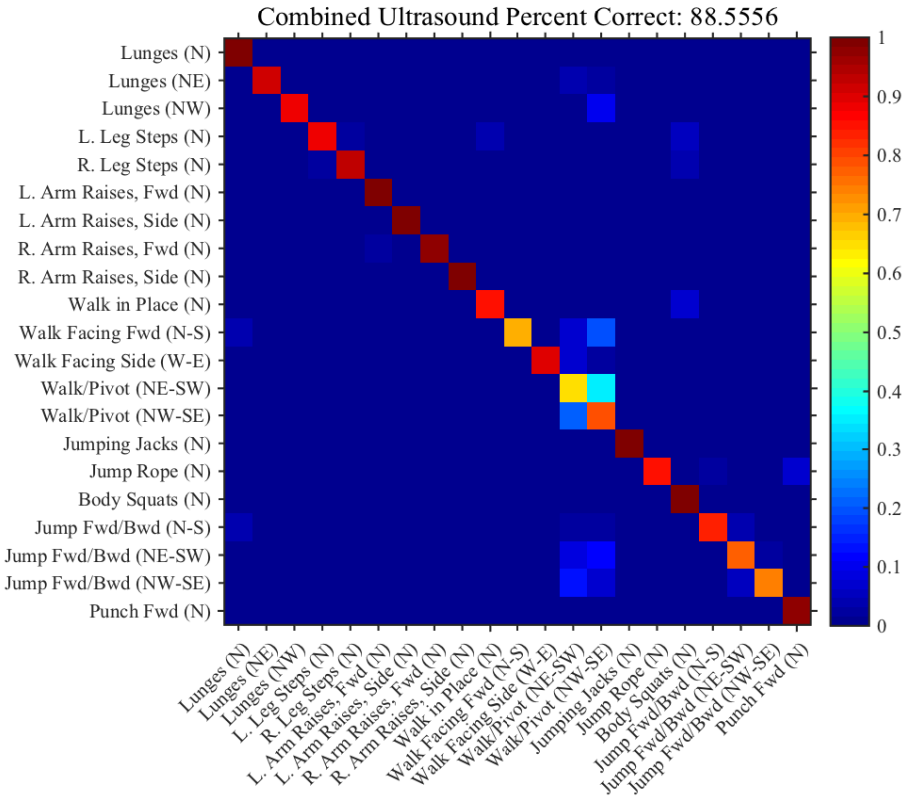

Fig. 22. Confusion matrix enumerating the action classification decisions resulting from combining the output of the three ultrasound models as a product of experts.

VII. Conclusion

Using a multimodal dataset that incorporates both visual data, which facilitates the accurate tracking of human movement, and active acoustic data, which captures the micro-Doppler modulations induced by the motion, we have developed algorithms for action recognition. The dataset consists of twenty-one actions and focuses on examples of orientational symmetry that a single active ultrasound sensor should have the most difficulty discriminating. The combined results from three independent ultrasound sensors are encouraging, and provide a foundation to explore the use of data from multiple viewpoints to resolve the orientational ambiguity in action recognition. Future lines of research are intended to explore the applicability of the sensor to real-life scenarios. In this sense, experiments will be developed to evaluate aspects such as the distance limits of the system, especially in outdoor conditions, and the effects on accuracy of the angle of incidence between the ultrasonic module and the target object. One key aspect here is the potential active control of the micro-Doppler sonar for interrogating the scene, as, unlike audio, which comes from all directions and without control, the sonar device can be activated intermittently and directed towards the desired objects.

Acknowledgments

The original development of the micro-Doppler units and the data acquisition system was supported by the EU ICT Grant (ICT-231168-SCANDLE) Acoustic SCene ANalysis for Detecting Living Entities. Data collection and algorithm development was supported by the ONR MURI (N000141010278) Figure-Ground Processing, Saliency and Guided Attention for Analysis of Large Natural Scenes. Further support was provided by an NSF grant (INSPIRE SMA 1248056) through the Telluride Workshop on Neuromorphic Cognition Engineering and by the NSF grant (SCH-INT 1344772). Dan Mendat was supported by a Johns Hopkins University Applied Physics Laboratory Graduate Student Fellowship.

References

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *Computing Surveys*, vol. 43, no. 3, Apr. 2011.

[2] P. Turaga, R. Chellapa, V. S. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[3] T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, Nov. 2006.

[4] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, 2004, pp. 32–36.

[5] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, Jul. 2013.

[6] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *arXiv.org*, Dec. 2012.

[7] S. Oh, A. Hoogs, A. Perera, and N. Cuntoor, "A large-scale benchmark dataset for event recognition in surveillance video," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[8] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandréa, C.-E. Bichot, C. Garcia, and B. Sankur, "The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition," Tech. Rep. LIRIS-RR-2012-004, 2012.

[9] Y. G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. (2014, Oct.) THUMOS Challenge 2014. [Online]. Available: http://crcv.ucf.edu/THUMOS14/

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Jun. 2014.

[11] J. Eklof, "Vision in echolocating bats," Ph.D. dissertation, Ph.D. Dissertation, Gotenborg University, Apr. 2003.

[12] C. F. Moss and A. Srlykke, "Probing the natural scene by echolocation in bats," *Frontiers in Behavioral Neuroscience*, pp. 1–16, 2010.

[13] G. von der Emde and H.-U. Schnitzler, "Classification of Insects by Echolocating Greater Horseshoe Bats," *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, vol. 167, no. 3, pp. 423–430, 1990.

[14] K. Koselj, H.-U. Schnitzler, and B. M. Siemers, "Horseshoe bats make adaptive prey-selection decisions, informed by echo cues," *Proceedings of the Royal Society B: Biological Sciences*, vol. 278, no. 1721, pp. 3034–3041, Sep. 2011.

[15] C. Doppler, "Über das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels (English Translation)," *Proceedings of the Royal Bohemian Society of Sciences*, vol. 2, pp. 465–482, 1842.

[16] B. Ballot, "Akustische Versuche auf der Niederländischen Eisenbahn, nebst gelegentlichen Bemerkungen zur Theorie des Hrn. Prof. Doppler," *Annalen der Physik and Chemie*, vol. 11, pp. 321–351, 1845.

[17] V. Chen, *The Micro-Doppler Effect in Radar*, ser. Artech House Remote Sensing Library. Artech House, Dec. 2010.

[18] Z. Zhang, P. O. Pouliquen, A. M. Waxman, and A. G. Andreou, "Acoustic micro-Doppler radar for human gait imaging," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. EL110–3, Mar. 2007.

[19] K. Kalgaonkar and B. Raj, "Acoustic Doppler sonar for gait recognition," in *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007)*, Aug. 2007, pp. 27–32.

[20] V. C. Chen, F. Li, S. S. Ho, and H. Wechsler, "Micro-Doppler effect in radar: phenomenon, model, and simulation study," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 1, pp. 2–21, 2006.

[21] M. Otero, "Application of a continuous wave radar for human gait recognition," in *Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition XIV*, May 2005, pp. 538–548.

[22] J. M. Carcia-Rubia, O. Kilic, V. Dang, Q. Nguyen, and N. Tran, "Analysis of Moving Human Micro-Doppler Signature in Forest Environments," *Progress in Electromagnetics Research*, vol. 148, pp. 1–14, Jun. 2014.

[23] Z. Zhang and A. G. Andreou, "Human identification experiments using acoustic micro-Doppler signatures," in *Proceedings of the 3rd Argentine School of Micro-Nanoelectronics, Technology and Applications (EAMTA 2008)*, 2008, pp. 81–86.

[24] G. Garreau, C. M. Andreou, A. G. Andreou, J. Georgiou, S. Dura-Bernal, T. Wennekers, and S. L. Denham, "Gait-based person and gender recognition using micro-doppler signatures," in *Proceedings of the 2011 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2011, pp. 444–447.

[25] A. Balleri, K. Chetty, and K. Woodbridge, "Classification of personnel targets by acoustic micro-Doppler signatures," *IET Radar, Sonar & Navigation*, vol. 5, no. 9, p. 943, 2011.

[26] G. Garreau, N. Nicolaou, and J. Georgiou, "Individual classification through autoregressive modelling of micro-doppler signatures," in *Proceedings of the 2012 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2012, pp. 312–315.

[27] T. Thayaparan, L. Stankovic, and I. Djurovic, "Micro-Doppler-based target detection and feature extraction in indoor and outdoor environments," *Journal of the Franklin Institute*, vol. 345, no. 6, pp. 700–722, 2008.

[28] K. Kalgaonkar and B. Raj, "Ultrasonic Doppler sensor for speaker recognition," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4865–4868.

[29] ——, "One-handed gesture recognition using ultrasonic Doppler sonar," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 1889–1892.

[30] G. Garreau, N. Nicolaou, C. M. Andreou, C. D'Urbal, G. Stuarts, and J. Georgiou, "Computationally efficient classification of human transport mode using micro-doppler signatures," in *Proceedings of the 45th Annual Conference on Information Sciences and Systems (CISS)*, 2011, pp. 1–4.

[31] G. Blumrosen, B. Fishman, and Y. Yovel, "Noncontact Wideband Sonar for Human Activity Detection and Classification," *Sensors Journal*, 2014.

[32] S. Dura-Bernal, G. Garreau, J. Georgiou, A. G. Andreou, S. L. Denham, and T. Wennekers, "Multimodal integration of micro-Doppler sonar and auditory signals for behavior classification with convolutional networks," *International Journal of Neural Systems*, vol. 23, no. 5, pp. 1 350 021–1 1 350 021–15, 2013.

[33] T. S. Murray, D. R. Mendat, P. O. Pouliquen, and A. G. Andreou, "The Johns Hopkins University Multimodal Dataset for Human Action Recognition," in *Proceedings of SPIE: Radar Sensor Technology XIX; and Active and Passive Signatures VI*, May 2015, pp. 79–94.

[34] J. Georgiou, P. O. Pouliquen, A. S. Cassidy, G. Garreau, C. M. Andreou, G. Stuarts, C. d'Urbal, S. L. Denham, T. Wennekers, R. Mill, I. Winkler, T. M. Bohm, O. Szalardy, G. M. Klump, S. Jones, A. Bendixen, and A. G. Andreou, "A multimodal-corpus data collection system for cognitive acoustic scene analysis," in *Proceedings of the 45th Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2011, pp. 1–6.

[35] B. Freedman, A. Spunt, and Y. Ariell, "Distance-varying illumination and imaging technique for depth mapping," Patent, Jun., 2014.

[36] G. Yahav, G. Iddan, and D. Mandelboum, "3D imaging camera for gaming application," in *Consumer Electronics 2007*, 2007, pp. 1–2.

[37] L. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE Signal Processing Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[38] F. Jelinek, *Statistical methods for speech recognition*. The MIT Press, 1998.

[39] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proceedings of the 1997 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997.

[40] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, 1st ed. The MIT Press, Jul. 2009.

[41] K. P. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," Ph.D. dissertation, Ph.D. Dissertation, University of California Berkeley, 2002.

[42] K. Murphy, "A brief introduction to graphical models and Bayesian networks," *http://people.cs.ubc.ca/ murphyk/Bayes/bnintro.html*, 1998.

[43] K. P. Murphy, *Machine Learning: a Probabilistic Perspective*. MIT Press, Sep. 2013.

[44] S. J. Young, G. Evermann, M. J. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*. University of Cambridge, Mar. 2009, vol. ver 3.4.

[45] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Oakland, CA, USA., 1967, pp. 281–297.

[46] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[47] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[48] A. Coates and A. Y. Ng, "Learning Feature Representations with k-means," in *Neural Networks: Tricks of the Trade*. Springer Lectures in Computer Science, 2012, pp. 561–580.

[49] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[50] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[51] D. Moore and I. Essa, "Recognizing multitasked activities from video using stochastic context-free grammar," in *Proceedings of Eighteenth National Conference on Artificial Intelligence*. American Association for Artificial Intelligence, Jul. 2002.

[52] M. S. Ryoo and J. K. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," in *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[53] H. Kjellström, J. Romero, D. Martínez, and D. Kragic, "Simultaneous visual recognition of manipulation actions and manipulated objects," in *Proceedings of the 10th European Conference on Computer Vision (ECCV'08)*. Springer, 2008, pp. 336–349.
[54] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
[55] B. Laxton, J. Lim, and D. Kriegman, "Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video," in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.
[56] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*. IEEE, 2003, pp. 742–749.
[57] C. S. Pinhanez and A. F. Bobick, "Human action detection using PNF propagation of temporal constraints," in *Proceedings of the 1998 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998, pp. 898–904.
[58] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto, "Recognition of human gaits," in *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.
[59] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1932–1939.